The Fate of Empirical Economics When All Data are Private

John M. Abowd
Cornell University and U.S. Census Bureau*
Society of Government Economists
May 13, 2016

*This work was performed as a Cornell faculty member before I joined the U.S. Census Bureau in an executive capacity. No confidential data from any source were used to develop this talk.

First, a Live Demo!

- My assistants are now distributing sealed envelopes:

CHOOSE YOUR OWN ENVELOPE

DO NOT LET THE ASSISTANT CHOOSE

PLEASE DO NOT OPEN THE ENVELOPE!!!!!!!!!

Step 1

OPEN THE ENVELOPE CAREFULLY (IT IS RESEALABLE)

REMOVE THE SHEET OF PAPER THAT SAYS
“CIRCLE ONE”

LEAVE THE OTHER SHEET OF PAPER IN THE ENVELOPE

Step 2

WITHOUT TAKING THE OTHER SHEET OF PAPER OUT OF THE ENVELOPE:

READ THE QUESTION INSIDE YOUR ENVELOPE

MEMORIZE YOUR ANSWER

RESEAL THE ENVELOPE WITH THE QUESTION INSIDE

Step 3

ANSWER THE QUESTION ON THE SHEET OF PAPER THAT SAYS

“CIRCLE ONE”

HAND YOUR ANSWER TO ONE OF THE ASSISTANTS

Step 4

WHILE THE ASSISTANTS ARE TABULATING THE DATA,

SHRED YOUR QUESTION ENVELOPE
IN ONE OF SHREDDERS IN THE ROOM

Now, Let's Analyze the Data

Parameter	*Value*		*Interpretation*
N	67		Sample size
Yes	28		Response to the survey question
No	39		Response to the survey question
beta_hat	41.8%		Raw percentage "Yes" estimate
Var[beta_hat]	0.0036		Sampling variance of raw proportion
StE[beta_hat]	6.0%		Standard error of raw percentage
Prec[beta_hat]	275		Sampling precision of raw proportion (inverse of sampling variance)
rho	0.5000		Probability that the sensitive question was asked
mu	0.5000		Probability of "Yes" on nonsensitive question
pi_hat	33.6%		Estimated percentage "Yes" to sensitive question
Var[pi_hat]	0.0145		Sampling variance of proportion "Yes" to sensitive question
StE[pi_hat]	12.1%		Standard error of percentage "Yes" to sensitive question
Prec[pi_hat]	69		Sampling precision of proportion "Yes" to sensitive question
Relative Precision	0.2500		Ratio of sampling precision of "Yes": sensitive question/raw question
Bayes Factor	3.0		
ln Bayes Factor	1.0986		

Randomized Response

As a survey technique:

- Warner, Stanley L. (1965) “Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias” Journal of the American Statistical Association 60, no. 309: 63–69, DOI: 10.2307/2283137.

As a privacy-preserving data analysis system

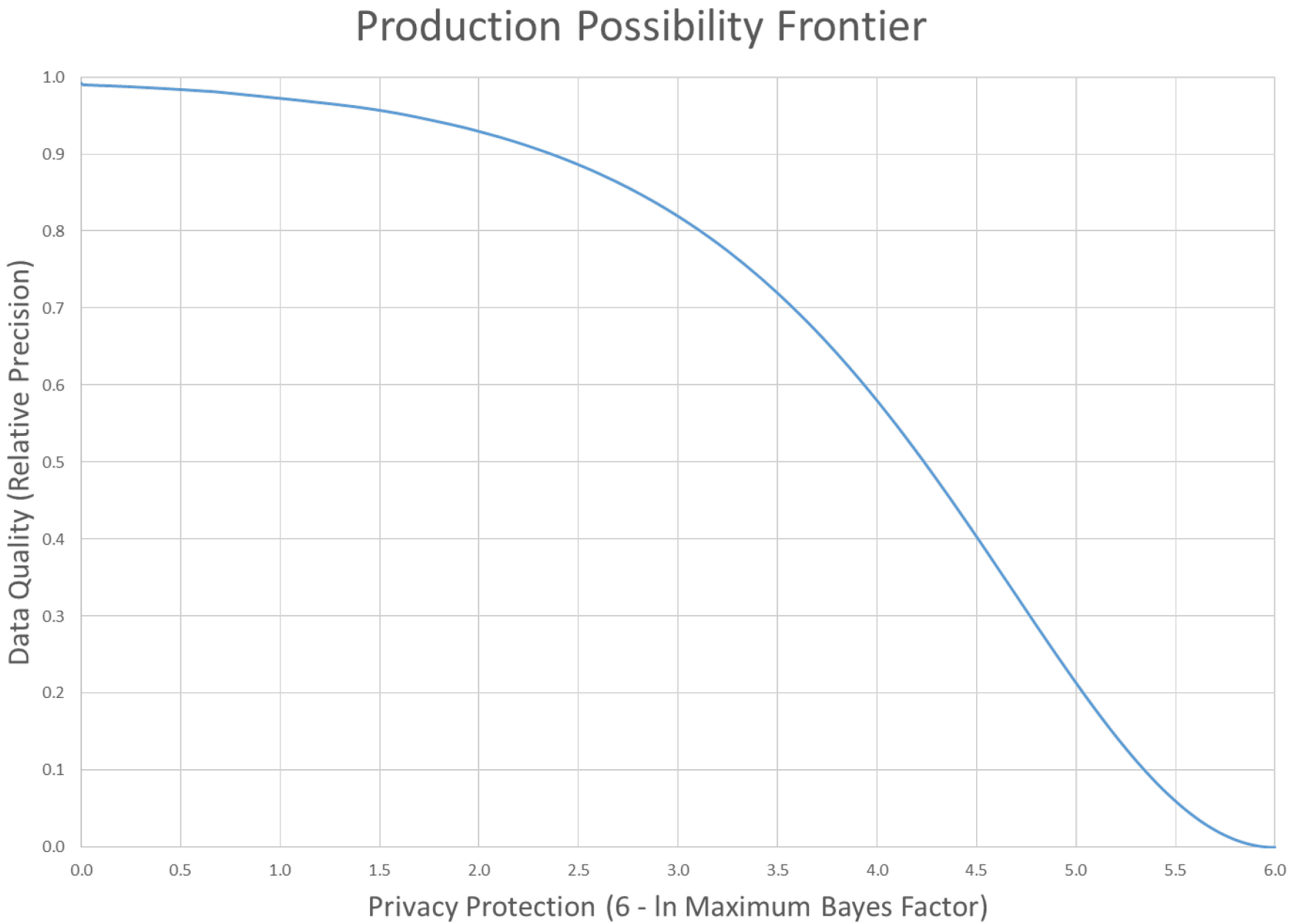
- Du and Zhan (2003) “Using Randomized Response Techniques for Privacy-Preserving Data Mining” SIGKDD ’03, August 24-27, 2003, Washington, DC, USA. DOI: 10.1145/956750.956810.
- Dwork and Roth (2014) “The Algorithmic Foundations of Differential Privacy.” Foundations and Trends in Theoretical Computer Science 9, nos. 3–4: 211–407.

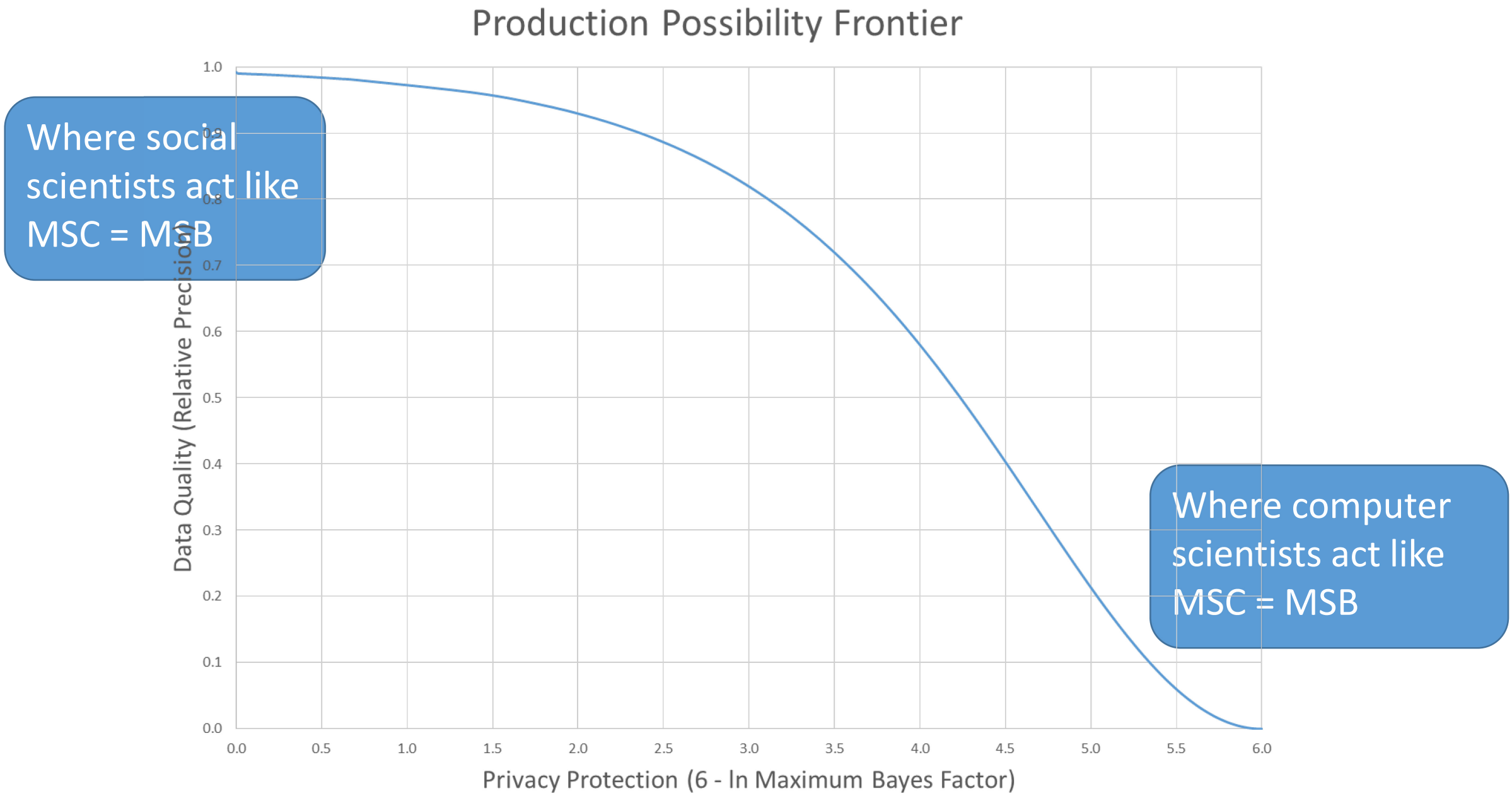
The Basic Economics

- Scientific data quality is a pure public good (non-rival, non-excludable)
- Quantifiable privacy protection is also a pure public good (or “bad,” when measured as “privacy loss”) when supplied using the methods I will discuss shortly
- Computer scientists have succeeded in providing feasible technology sets relating the public goods: data quality and privacy protection
- These technology sets generate a quantifiable production possibilities frontier between data quality and privacy protection

The Basic Economics II

- We can now estimate the marginal social cost of data quality as a function of privacy protection—a big step forward
- The CS models are silent (or, occasionally, just wrong) about how to choose a socially optimal location on the PPF because they ignore social preferences
- To solve the social choice problem, we need to understand how to quantify preferences for data quality v. privacy protection
- For this we use the Marginal Social Cost of data quality and the Marginal Social Benefit of data quality, both measured in terms of the required privacy loss


Production Possibility Frontier
Data Quality (Relative Precision)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
Privacy Protection (6 - ln Maximum Bayes Factor)


Production Possibility Frontier
Where social scientists act like MSC = MSB
Where computer scientists act like MSC = MSB
Data Quality (Relative Precision)
Privacy Protection (6 - ln Maximum Bayes Factor)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0

How Should We Measure MSB?

- Medical diagnosis example
- Consumer price index example
- Legislative apportionment example
- Generically: sum of all the marginal social benefits from every potential use
- Not: marginal social benefit of the highest-valued user (market solution)

Ideal Data Publication - Privacy Protection Systems

- To the maximum extent possible, *scientific analysis* should be performed on the *original confidential data*
- Publication of *statistical results* should respect a quantifiable *privacy-loss budget* constraint
- Data *publication algorithms* should provably *compose*
- Data *publication algorithms* should be provably *robust to arbitrary ancillary information*

Doing Data Analysis in This World

- Census Bureau already does this in some applications
  - OnTheMap
  - Survey of Income and Program Participation Synthetic Data
  - Synthetic Longitudinal Business Database
- Google does this
  - Randomized Aggregatable Privacy Preserving Ordinal Responses (tool for Cloud service providers to harvest browser data)
- Prototype systems allow medical record databases to do this
  - Privacy-preserving deep learning
  - Computational healthcare

Doing Data Analysis in This World

See:

- Abowd and Schmutte “Economic Analysis and Statistical Disclosure Limitation” Brookings Papers on Economic Activity (Spring 2015), http://www.brookings.edu/~/media/Projects/BPEA/Spring-2015-Revised/AbowdText.pdf?la=en
- Erlingsson, Pihur and Korolova “RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response” CCS’14, November 3–7, 2014, Scottsdale, Arizona, USA. ACM 978-1-4503-2957-6/14/11, http://dx.doi.org/10.1145/2660267.2660348
- Shokri and Shmatikov “Privacy-Preserving Deep Learning” CCS’15, October 12–16, 2015, Denver, Colorado, USA. ACM 978-1-4503-3832-5/15/10, http://dx.doi.org/10.1145/2810103.2813687

Thank you!

Contacts:

john.abowd@cornell.edu